STYLEFORMER

A Convolution-Free Style Image Generator based on Transformer And Stylegan2.

Sapienza
University of Rome

Fabio Caputo

Weihao Peng

STRONG STYLE GENERATOR CONVOLUTION - FREE

GAN's (Generative Adversarial Networks) models are living a huge success since their introduction in 2014. Nowadays resolution and quality of the generated images increased a lot, despite that, there is something that doesn't changed: the consideration of convolutional backbones as fundamental to achieve high-resolution images and a stable training. In this work, we have tried to implement a strong, and light, style-based generator designed with as a convolution-free structure based on NPL technologies like Transformer and Attention.

TECH OVERVIEW

How GAN works and why we are trying to use NPL technologies to generate images.

- First GAN model was introduced by Ian Goodfellow in 2014.

- StyleGAN is a progressive growing GAN architecture, able to synthesizing high resolution and quality images with incremental growing of discriminator and generator.

- That model shows some problems in generation, StyleGAN2 addressed most them using skip connection and replacing AdaIN with a statistic-based approach.

- Some problem remains, shortcomings derived using a convolutional network such locality problem led to a difficult capture of the global features.

STYLE
GENERATIVE
ADVERSARIAL
NETWORK

ALL YOU NEED IS TRANSFORMER

- *"The first transduction model **relying entirely on self-attention** to compute representations of its input and output without using sequence-aligned RNNs or convolution"*

- Designed for NPL, recently is rising as an alternative to convolution operation in the computer vision field.

- Based on **attention**, a mechanism that mimic the cognitive attention focusing on small but significative details of an image, a token or any other significative data.

- Stacking attention and combining them with feed-forward layers, we can form encoders (**self-attention**).

ALL
YOU NEED IS
TRANSFORMER

- Solves the difficult to capture long-range dependency without stacking multiple layers.

- Indeed, using self-attention we are able to capture long-range dependency and understand global features efficiently.

- Using Linformer we can address Transformer expensive cost while dealing with high-resolution images.

ARCHITECTURE OVERVIEW

How Styleformer is built, which are the main components and how it works

ARCHITECTURE
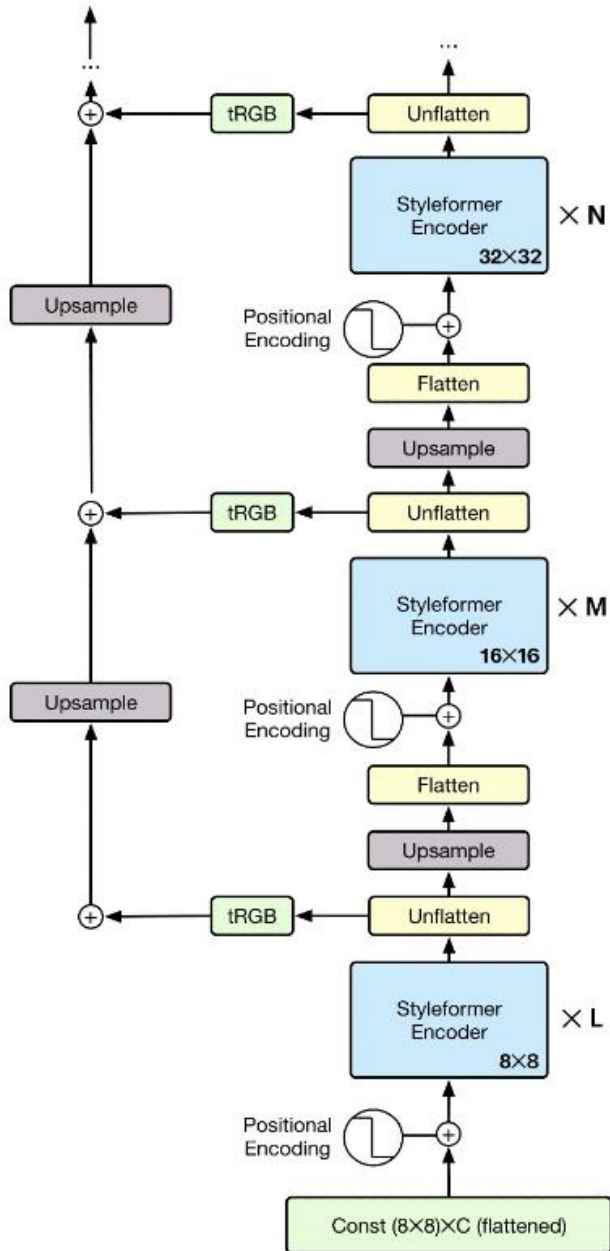
- Styleformer generator is conditioned on a learnable constant input and combined with a learnable positional encoding.

- The constant input(8x8) is flattened(64) to enter the Transformer-based encoder, then the Styleformer encoder.

- Each resolution passes through several encoder blocks. ⟶ Bilinear upsample operation. ⟶ Adding positional encoding.

- Repeat until reaches the target image resolution.

ARCHITECTURE
ENCODER BLOCK

- We need a Transformer-based generator

- We need a style modulation and demodulation methods

Attention mechanism can be seen as built in two steps

1. preparation module: compute Query, Key and Value

2. main module: attention operation is applied

COMPUTING ATTENTION

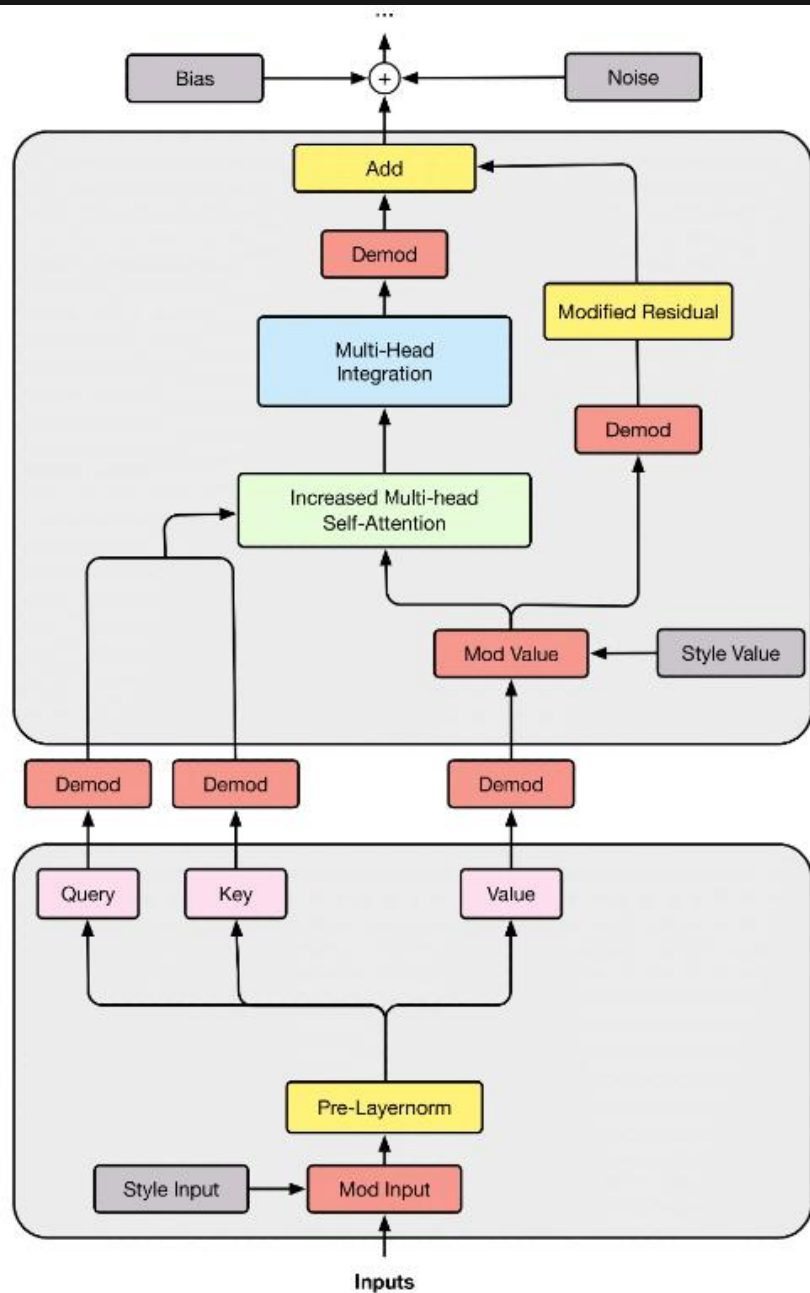- Preparation module is a module that creates Quey, Key and Value to conduct attention:

$$Q_i = XW_i^Q, \ K_i = XW_i^K, \ V_i = XW_i^V$$

- Main module is a module that performs attention operation:

$$A_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}}),$$

$$head_i = A_i V_i,$$

$$Multihead(Q, K, V) = Concat(head_1, \ldots, head_k)W^O$$

ARCHITECTURE
FOCUS ON ENCODER BLOCK

- Pre-Layer Normalization procedure

- Demodulation of Query, Key and Value

- Core of Self-Attention: production of the attention map and the weighted sum of V with attention map itself

- Increased the number of head of multi-head attention ⟶ the created attention map will be different for each head

- Further demodulation

- Add bias and noise at end of each blocks

CONCLUSIONS

HOW GOOD STYLEFORM IS?

CIFAR-10	FID	IS
PROGRESSIVE GAN	15.52	8.80
TRANSGAN	9.26	9.02
STYLEGAN2	2.92	9.83
STYLEFORMER	2.82	10.0

- CIFAR-10 is widely used as a benchmark dataset. They used 50K images(32x32) at the training set, without using label.

- We have used two of the most known metrics for image evaluation:

- Fréchet Inception Distance, known as FID, is a method for comparing the statistics of two distributions by computing the distance between them

- Inception score, known as IS, is a popular metric for judging the image outputs of GAN which measure of how realistic a GAN's output is.

CONCLUSIONS

- Training and image generation.

- Assembling a custom and topic-based dataset to train on it.

- Assembling generate images to achieve a morph effect.

- Exploring latent vectors direction performing fine tune.

CONCLUSIONS

WHAT'S NEXT?

- Test Styleformer with high end resources to have better train results in less time.

- Reduce the computational cost to test with higher-resolution images.

- Redesign also the discriminator to use Transformer.

- Porting this project to Stylegan3.

THANK YOU!

Styleformer Implementation

Google Colab Notebook

Fabio Caputo

Weihao Peng